

Microscopic Approaches for the Discovery of Web Communities (extended abstract)

Tsuyoshi Murata

National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan
tmurata@nii.ac.jp, <http://research.nii.ac.jp/~tmurata>

1 Introduction

The Web can be regarded as a graph if we regard each Web page as a vertex and each hyperlink as an edge. There are several goals for Web structure mining based on the graph structure of hyperlinks, such as ranking important Web pages [4][12], discovery of Web communities [3][5], analysis of the Web graph from macroscopic point of view [2], and modeling and simulating the process of Web graph generation [1]. Among these, discovery of Web communities (clusters of related Web pages whose hyperlinks are densely connected) is important in order to assist users' information retrieval from the Web. This paper briefly introduces our microscopic approaches for the discovery of Web communities.

2 Discovery of Web communities using a search engine

A search engine can be regarded as a resource for Web data acquisition. The author proposed a method for discovering Web communities from the data acquired from a search engine [6][7]. Our method is similar to Kumar's method [5] since both perform search bipartite graph structures. However, they are different in the following points:

1. Search of bipartite graphs from partial Web data without using Web snapshot data

Previous approaches of Web community discovery require relatively large-scale Web snapshot data. However, collecting Web data and maintaining them is not an easy task. It is pointed out that the difference between Web snapshot data used for mining and actual Web data may cause the discovery of outdated Web communities [5]. Major search engines contain much updated Web data and they can be used for Web data acquisition in order to achieve relatively new Web communities. Some of the search engines allow users to access contained data, such as Google API.

2. Acquisition of backlinks from a search engine in order to follow hyperlink backward

Although most users use search engines in order to find Web pages about some keywords, a search engine enables us to follow hyperlinks backward. By attaching some option (such as "link:") to input URL, Web pages that contain hyperlinks to

input URLs can be searched, which are called backlinks. Since hyperlinks to related Web pages often co-occur, backlink search enables us to find related Web pages.

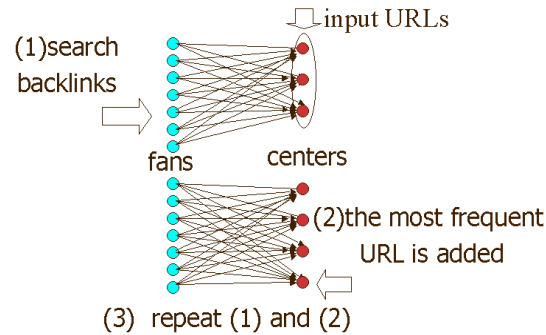


Fig. 1. Outline of our method for discovering Web communities

Fig.1 shows the outline of our method for Web community discovery. Our goal is to discover a bipartite graph containing some given URLs. At first, some URLs regarding specific topic (such as baseball or Macintosh) are given as initial centers, and fans which co-refer all of the centers are searched by backlink search on a search engine (step 1). HTML files of the searched fans are acquired through the internet, and all the hyperlinks contained in the files are extracted. The hyperlinks are sorted in the order of frequency. Since hyperlinks to related Web pages often co-occur, the top-ranking hyperlink of the sorted result is expected to point to a page whose contents are closely related to the contents of centers. Therefore, the URL of the page is added as a new member of centers (step 2). By using newly generated centers, the above steps are repeated in order to find more centers (step 3). Although this method is quite simple, it succeeds in discovering many related Web pages. Experimental results show that 19.8 related centers are actually discovered from given 5 seed URLs on average [7].

3 Microscopic approaches for discovery of Web communities

There are two main approaches for the discovery of Web communities; 1) search of fixed-size graph structure from Web snapshot data [5], and 2) decomposition of given Web graph into densely connected components [3]. Our method described above performs search of bipartite graph by using local Web data. Its results give insights for microscopic topological properties of Web graph structure, which are important for analyzing the effectiveness of propagation algorithms. Other attempts by the author are purifying Web communities [8], visualizing the structure of Web communities [9], and detecting boundaries of Web communities from positive and negative examples [10].

References

1. Barabasi, A.-L., "LINKED – The New Science of Networks", Perseus Publishing, 2002.
2. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: "Graph Structure in the Web: Experiments and models", Proc. of the 9th WWW Conference, pp.309-320, 2000.
3. G. W. Flake, S. Lawrence, C. L. Giles, F. M. Coetzee: "Self-Organization and Identification of Web Communities", IEEE Computer, Vol.35, No.3, pp.66-71, 2002.
4. J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins: "The Web as a Graph: Measurements, Models, and Methods", Proc. of COCOON'99, Lecture Notes in Computer Science 1627, pp.1-17, 1999.
5. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins: "Trawling the Web for Emerging Cyber-Communities", Proc. of the 8th WWW Conference, 1999.
6. T. Murata: "Discovery of Web Communities Based on the Co-occurrence of References", Proc. of the Third International Conf. on Discovery Science (DS2000), Lecture Notes in Artificial Intelligence 1967, pp.65-75, Springer, 2000.
7. T. Murata: "Finding Related Web Pages Based on Connectivity Information from a Search Engine", Poster Proc. of 10th WWW conference, pp.18-19, 2001.
8. T. Murata: "A Method for Discovering Purified Web Communities", Proc. of the Fourth International Conf. on Discovery Science (DS2001), Lecture Notes in Artificial Intelligence 2226, pp.282-289, Springer, 2001
9. T. Murata: "Visualizing the Structure of Web Communities Based on Data Acquired from a Search Engine", IEEE Transactions on Industrial Electronics, Vol. 50, No. 5, pp.860-866, 2003.
10. T. Murata: "Discovery of Web Communities from Positive and Negative Examples", Proc. of the Sixth International Conference on Discovery Science (DS2003), Lecture Notes in Artificial Intelligence 2843, pp.365-372, Springer, 2003.
11. T. Murata: "Graph Mining Approaches for the Discovery of Web Communities", Proc. of the First International Workshop on Mining Graphs, Trees and Sequences (MGTS-2003), pp.79-82, 2003.
12. L. Page, S. Brin, R. Motwani, T. Winograd.: "The PageRank Citation Ranking: Bringing Order to the Web", Online manuscript, [http://www-db.stanford.edu/~backrub/pagerank sub.ps](http://www-db.stanford.edu/~backrub/pagerank.sub.ps), 1998.